Translational efficiency across healthy and tumor tissues is proliferation-related

Xavier Hernandez-Alias[1], Hannah Benisty[1], Martin H. Schaefer[1,2,]*, Luis Serrano[1,3,4,]*

[1]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, Barcelona 08003, Spain

[2]IEO European Institute of Oncology IRCCS, Department of Experimental Oncology, Via Adamello 16, Milan 20139, Italy

[3]Universitat Pompeu Fabra (UPF), Barcelona 08002, Spain

[4]ICREA, Pg. Lluís Companys 23, Barcelona 08010, Spain

*Corresponding authors: martin.schaefer@ieo.it (M.H.S.), luis.serrano@crg.eu (L.S.)

ABSTRACT

Background: Different tissues express genes with particular codon usage and anticodon tRNA repertoires. However, the codon-anticodon co-adaptation in humans is not completely understood, as well as its effect on tissue-specific protein levels.

Results: We first validated the accuracy of small RNA-seq for tRNA quantification across five human cell lines. We then analyzed tRNA expression in more than 8000 tumor samples from TCGA, together with their paired mRNA-seq and proteomics data, to determine the Relative Translation Efficiency. We thereby elucidate that the dynamic adaptation of the tRNA pool is largely related to the proliferative state across tissues, which determines tissue-specific translation efficiency. Furthermore, the aberrant translational efficiency of ProCCA and GlyGGT in cancer, among other codons, which is partly regulated by the tRNA gene copy numbers and their promoter DNA methylation, is associated with poor patient survival.

Conclusions: The distribution of tissue-specific tRNA pools over the whole cellular translatome affects the subsequent translational efficiency, which functionally determines a condition-specific expression program in tissues both in healthy and tumor states.

KEYWORDS

tRNA, translation, The Cancer Genome Atlas, tissue, regulation, miRNA, codon usage

BACKGROUND

In the light of the genetic code, multiple 3-letter combinations of nucleotides in the mRNA can give rise to the same amino acid, which are known as synonymous codons. However, despite the homology at the protein level, these different codons are recognized distinctly by the transcriptional and translational machineries (1,2), and ultimately cause changes at multiple levels of gene expression. Therefore, the non-uniform abundance of synonymous codons across different tissues and among distinct functional gene sets has been proposed as an adaptive mechanism of gene expression regulation (3), particularly linked to the proliferative state (4). Nevertheless, in human, it is still under debate whether the efficiency of gene expression is the main selective pressure driving the evolution of genomic codon usage (5).

The 61 amino-acid-coding codons need to be recognized by 46 different tRNA isoacceptors distributed across 428 Pol-III-transcribed tRNA genes (6), thus requiring wobble interactions (non-Watson-Crick base pairing). This complexity of the tRNA repertoire is further enhanced by an average of 11-13 base modifications per tRNA and all possible combinations thereof (7). The underlying mechanisms regulating tRNA gene expression and modification are far from resolved (8,9). However, it has been established that different conditions and tissues showcase distinct tRNA abundances (4,10) and codon usages (3,11).

In order to understand such changes in codon-anticodon co-adaptation, orthogonal datasets of gene expression including tRNA quantification are required, which needs to overcome the challenges of strong secondary structures and abundant chemical modifications. Recent technological developments have paved the way for sensitive high-throughput tRNA sequencing across tissues and conditions (12,13). Aside from these methods and despite the lower coverage, tRNA reads can also be detected from generic small RNA-seq datasets

(14–18). In this context, The Cancer Genome Atlas (TCGA) has been recently used to investigate the alteration of tRNA gene expression and translational machinery in cancer, which may play a role in driving aberrant translation (19,20).

To validate the use of small RNA-seq for tRNA quantification, we first compare tRNA levels determined in HEK293 by well-established tRNA sequencing methods (Hydro-tRNAseq and demethylase-tRNA-seq) (12,13,21), with those obtained by small RNA-seq. Then we quantify the tRNA repertoire of five cell lines using Hydro-tRNAseq and perform small RNA-seq in parallel. Comparison of the tRNA abundance obtained by both approaches shows that it is possible to accurately estimate relative tRNA abundance of cells and tissues using small RNA-seq. Furthermore, we show that both types of quantification are informative enough to distinguish between the five analyzed human cell lines covering multiple tissue types. In consequence, we apply a tRNA-specific computational pipeline to re-analyze 8,534 small RNA-seq datasets from TCGA (22). We find that the tissue-specificity of tRNA expression is largely proliferation-related, even within healthy tissues. The tRNA quantification of TCGA samples enables their comparison with paired and publicly available mRNA-seq, proteomic, DNA methylation and copy number data, which underscores the role of tRNAs in globally controlling a condition-specific translational program. We discover multiple codons, including ProCCA and GlyGGT, whose translational efficiency is compromised and leads to poor prognosis in cancer. Finally, promoter DNA methylation and tRNA gene copy number arise as two regulatory mechanisms controlling tRNA gene expression in cancer.

RESULTS

tRNA quantification and modifications from small RNA-seq data

In order to test how accurately we can extract tRNA abundance information contained in small RNA sequencing data, we re-analyze four publicly-available datasets of the cell line HEK293 (18,23,24). In contrast to previous studies analyzing tRNA expression from small RNA-seq data (19,20), we use a computational pipeline specifically developed for the accurate mapping of tRNA reads (16) in order to quantify all different isoacceptor species (Figure 1A, see Methods). To validate the accuracy of these small RNA-seq quantifications, we retrieve three datasets of well-established tRNA sequencing methods (Hydro-tRNAseq and demethylase-tRNA-seq) applied to the same cell type (12,13,21), which autocorrelate in the range of 0.72-0.79 among themselves (Table S1). In comparison, our four HEK293 small RNA-seq quantifications show an average Spearman correlation against these three conventional datasets of 0.68. Compared to the anticodon Spearman correlation of 0.61 as published by Zhang et al. (19), our tRNA-specific mapping pipeline performs better than the previously published protocol, since all our correlations lie over that value.

Further than correlating small RNA-seq data with conventional tRNA-seq datasets, we analyze whether small RNA-seq quantifications are informative enough to distinguish between different human cell lines covering multiple tissue types. We therefore apply both small RNA-seq and Hydro-tRNAseq to HEK293 (kidney), HCT116 (colon), HeLa (cervix), MDA-MB-231 (breast), and BJ fibroblasts. However, given the high variability between replicates of MDA-MB-231 Hydro-tRNAseq quantifications, this cell line was excluded from further analyses (Table S2). First, the correlations between the two methods of identical samples and computational mapping pipeline range between 0.93 and 0.96 for all cell lines. tRNA quantifications from both protocols are compared and significantly higher Spearman correlations are obtained within matching samples versus mismatching cell lines (Figure 1B). In consequence, we demonstrate that small RNA-seq quantifications of sample-specific tRNA profiles show a good agreement with Hydro-tRNAseq.

We also detect tRNA base modifications in both protocols by nucleotide variant calling, as described in Hoffmann et al. (16). In all cases, considering the modifications that are detected in all three replicates, Hydro-tRNAseq datasets identify a larger number of modifications than small RNA-seq, as expected by the more uniform and deeper coverage of this method (Table S2). Furthermore, we detect a significant enrichment of the Hydro-tRNAseq modifications in the small RNA-seq data (p < 1e-16, Fisher test), indicating that the latter contains also information on tRNA modifications (Figure 1C).

Taken together, these observations demonstrate the applicability of small RNA-seq data for the quantification tRNAs and their modifications. We therefore apply the same computational pipeline to all healthy and primary tumor small RNA-seq samples from 23 cancer types of The Cancer Genome Atlas (TCGA), which consists of 8,605 samples distributed among 17 different human tissues (Figure 1D, number of samples and their abbreviations in Table S3).

Figure 1. tRNA quantification and modifications from small RNA-seq data. (A) Schematic pipeline for accurate mapping of tRNA reads. (B) Correlations between tRNA quantifications by small RNA-seq and Hydro-tRNAseq of matching (correlations within the same cell line) versus non-matching (different cell lines) samples. The p-value corresponds to a one-tailed Wilcoxon rank-sum test, with $n_{matching}$= 9 and $n_{non-matching}$= 72. (C) Overlap of the detected tRNA modifications upon variant calling by both methods. (D) The TCGA network contains small RNA-seq data alongside mRNA-seq, DNA methylation arrays, non-targeted proteomics, and copy number alteration quantification comprising 17 tissues.

Proliferation is the major driver of tissue-specificity in tRNAs

To determine the tissue-specificity of tRNAs in physiological conditions, the tRNA levels of all 675 healthy samples in TCGA tissues are analyzed. The isoacceptor abundances show a significant tissue-specificity for all 46 annotated anticodons (q<0.05, FDR-corrected Kruskal-Wallis test grouped by cancer types). Such differences between tissues are also observed by hierarchical clustering of the median expression between all groups (Figure 2A). Furthermore, healthy samples from cancer types originating from the same tissue tend to cluster together: READ and COAD from the gut; KIRC, KIRP and KICH from the kidney; LUAD and LUSC from the lung; UCEC and CESC from the uterus; LIHC and CHOL from the liver (refer to Table S3 for full cancer names). On the other hand, in terms of anticodon abundances, three main subgroups of tRNAs with low, medium and high expression can be distinguished across all cancer types (Figure 2A).

Regarding codon usage, a measure of tRNA abundance taking into account the relative contribution of each tRNA anticodon among the set of codons of a certain amino acid is the Relative Anticodon Abundance. From this perspective, a principal component analysis (PCA) of the healthy control samples in TCGA also shows clear differences between tissues (Figure 2B). Moreover, our first component, which explains 18.5% of the variance, correlates positively with the proliferation marker Ki67 ($R_{spearman}$= 0.45) (25). To further interrogate the biological functions related to the variability of anticodon abundances between samples, we compute the correlation of the whole mRNA-seq transcriptome against the first PCA component, and analyze it by Gene Set Enrichment Analysis (GSEA). As a result, the top correlating genes are enriched in proliferation and immune cell activation, while the lowest correlations belong to genes related with oxidative metabolism and respiration (Figure 2C, Table S4). This confirms, as has been previously suggested (4), that there is a proliferative tRNA expression program.

Overall, we observe patterns of tissue-specific tRNA expression in TCGA healthy samples. Furthermore, our analysis identifies the proliferative state of tissues as the major biological function driving the variability on tRNA abundances.

Figure 2. Proliferation is the major driver of tissue-specificity in tRNAs. (A) Medians of square-root-normalized tRNA abundances across all TCGA tissues. The color of the tissue labels correspond to the average Ki67 expression. (B) Principal Component Analysis (PCA) of the Relative Anticodon Abundances (RAA) of TCGA, where the color scale corresponds to the mean tissue expression of Ki67. The Spearman correlations of Ki67 with the components are shown, as well as the samples of most extreme tissues. (C) Top positive and negative GO terms upon Gene Set Enrichment Analysis (GSEA) of the correlations of the first PCA component against all genes. Refer to Supplementary Table 3 for full cancer type names.

tRNA repertoires determine tissue-specific translational efficiency

Given that different tissues express distinct tRNA repertoires, we wondered whether they could have an effect in protein translation. In this context, and based on previous studies underscoring the global control role of codon usage as a competition for a limited tRNA pool (26–28), we define the Relative Translation Efficiency (RTE) as the balance between the supply (i.e. the anticodon tRNA abundances) and demand (i.e. the weighted codon usage based on the mRNA levels) for each of the 60 codons (excluding methionine and Stop codons). Furthermore, we normalize both the codon and anticodon abundances within each amino acid family (i.e. relative to the most abundant synonymous codon/anticodon), in order to remove the effect of amino acid biases and get a cleaner measure of codon optimality (29).

To validate the suitability of RTE in determining the translational efficiency, we correlate the RTE value of all proteins against the available proteomics data of paired TCGA samples (30,31), which includes breast and colorectal tissues (tumor only, as no healthy samples are available). Although the correlation is poor (but significant), both the protein abundances and the protein-to-mRNA ratios correlate significantly better with RTE than with the classical tRNA Adaptation Index [tAI] (32,33) or with a relative tAI with normalized weights within each amino acid family, which do not consider the mRNA codon demand (Figure 3A). Furthermore, the correlation of RTE with protein-to-mRNA ratio is slightly but significantly higher than with protein levels alone, which indicates that the first is a better proxy for the process of translation.

Next, we calculate the RTE for the 620 healthy samples for which both tRNA abundances and mRNA levels are available. When analyzing the tissue medians of RTE weights per each codon (RTEw), we observe that most codons are optimally balanced (RTEw =1), while 12.4% and 23.6% of codons are favored (RTEw >2) and disfavored (RTEw <0.5) respectively. The tissue clustering again shows that healthy samples of cancer types from the same tissue have similar RTEw profiles, which separates two major clusters of mostly high-Ki67 and low-Ki67 tissues (Figure S1).

In order to identify the codons contributing most to the differences between tissues, we compute a bidimensional PCA across all samples and RTEw (Figure 3B). Both the first and second components significantly correlate with the proliferation marker Ki67 (0.4 and 0.35; see Figure 2B). In agreement with the proliferation- and differentiation-related codons of Gingold et al. (4), such proliferative pattern is similarly reproduced by the codons contributing to the first PCA component, which has the strongest association to proliferation (Figure 3B). Further, similarly to the tRNA abundances (Figure 2B), a GSEA of correlating genes with the first component highlights the link with proliferation-related terms (Table S5). On the other hand, the first component also clearly separates codons based on the GC content of the third codon base, which has recently been associated with differentiation (high in nnC/G codons) versus self-renewal functions (high in nnA/T) (34), as well as with proliferative transcriptomes (35).

The previous analyses support the idea of proliferation-related tRNAs driving changes in translational efficiencies. In that case, we expect that the two most extreme tissues in terms of proliferation (brain and gut, excluding thymus for its low number of samples) differ in the optimization of proliferation-related proteins. As such, we compute the average RTEw for these two tissues, analyze the subsequent RTE score for each protein, and perform a GSEA of the differential RTE per protein. Consistent with our hypothesis, the results indicate that gut-optimized proteins are enriched in translation, DNA replication and protein localization, whereas brain-optimized proteins are related to phospholipid production and neural function (Figure 3C, Table S6). Taken together, this result confirms that the tRNA-dependent translational efficiency is optimized for the translation of tissue-specific genes, particularly in function of the proliferation state.

Figure 3. tRNA repertoires determine tissue-specific translational efficiency. (A) Three metrics of translation efficiency (the classical tAI, a relative tAI with normalized weights within each amino acid family, and the Relative Translation Efficiency described in this article) are Spearman correlated against two proxies of translation (protein abundance and protein-to-mRNA ratio) for all samples for which proteomics data is available (BRCA, COAD and READ). Statistical differences are determined by sample-paired two-tailed Wilcoxon rank-sum test. (B) Principal Component Analysis (PCA) of the RTEw of TCGA, where the color scale corresponds to the mean tissue expression of Ki67. The Spearman correlations of Ki67 with the components are shown, as well as the samples of most extreme tissues. On the right, the top and bottom proliferation- and differentiation-related codons, as defined by Gingold et al. (2014), ordered by their contribution to the first PCA component. (C) GSEA of the differential RTE between extreme tissues ($\Delta RTE = RTE_{Colorectal} - RTE_{Brain}$), showing the top five GO terms with high (left) and low (right) RTE in colorectal versus glial tissues. Refer to Supplementary Table 3 for full cancer type names.

Aberrant translational efficiencies drive tumor progression

Given that proliferation is a major determinant of translational efficiency in healthy tissues, its importance could be extrapolated to pathological conditions such as cancer. In fact, aberrant expression of tRNAs and codon usage have been broadly related with tumorigenesis and cancer progression (19,20,36,37). We therefore investigate 22 cancer types from TCGA in order to determine which codons are translationally compromised in disease.

Similar to the analysis performed on the healthy tissues, we quantify all tRNA abundances of TCGA primary tumor samples (Figure S2) and determine their corresponding translational efficiencies using the RTE metric. By analyzing the differential RTEw between normal and tumor samples, we observe many significant differences in all 60 codons across the 22 cancer types (Figure 4A). Among the most consistent changes, the ProCCA codon is significantly more favored in tumors for 8 out of 10 cancer types, while the ProCCG is disfavored in 14 out of 16 cancers (Figure 4B). In the case of glycine, translation appears more efficient for GlyGGT in healthy samples (13/13), whereas tumor mostly favors GlyGGC (9/12) and GlyGGG (7/9).

In terms of patient survival, we divide the TCGA patients in two groups based on their low or high tumor RTEw and analyze their survival probability (Figure 4C, Table S7). Among others, and consistent with the previous analysis, high translational efficiency weights of ProCCA are associated with poor prognosis in kidney renal clear cell carcinoma and kidney renal papillary cell carcinoma. Proline limitation in clear cell renal cell carcinoma has been shown to compromise CCA-decoding tRNAPro aminoacylation, leading to reduced tumor growth (38). In contrast, high RTEw of GlyGGT and ValGTC lead to longer survival in kidney chromophobe and head and neck squamous cell carcinoma, respectively.

To determine the impact of aberrant translational efficiencies in regulating an oncogenic translation program, we calculate the differential RTE for the whole genome based on the average RTEw of healthy and tumor samples in kidney renal clear cell carcinoma, since it is the cancer type with the most RTEw differences. The GSEA of the resulting ΔRTE score indicates that cancer RTEw enhance the translation of proteins related to DNA replication and gene expression, whereas the healthy kidney samples favor development and differentiation processes (Table S8). As the RTEw of the ProCCA is specifically disturbed in cancer, we also interrogate how this codon is distributed along the genome. We therefore perform a GSEA on the relative codon usage of ProCCA, which shows that DNA replication and cell cycle functions lie among the most CCA-enriched genes, while morphogenesis and differentiation terms are CCA-depleted (Table S9). Together with the low-proliferative state of kidney (Figure 2B), the over-efficiency of a proliferation-related codon in this tissue can thus perturb its cellular RTE.

Overall, we detect differences at the level of RTEw between tumor and healthy tissues, which show a functional relevance to the disease state. Therefore, while the differential expression of tRNAs in TCGA had been already discussed elsewhere (19,20), we could here elucidate their oncogenic effect in translational efficiency. In particular, ProCCA appears as an interesting codon candidate in favoring tumor progression, which we had also detected in healthy tissues to be associated with proliferation (Figure 3B, Table S5).

Figure 4. Aberrant translational efficiencies drive tumor progression. (A) Differential RTEw between healthy and tumor samples across 22 cancer types, as measured by $\log_2(\mathrm{RTEw}_{\mathrm{Tumor}}/\mathrm{RTEw}_{\mathrm{Healthy}})$. Only significant differences are colored, which are determined using a two-tailed Wilcoxon rank-sum test and corrected for multiple testing by FDR. (B) Boxplot of the RTEw of ProCCA and AlaGCG codons across TCGA cancer types. (C) Survival curves for the previous codons in KIRC, KIRP and BLCA patients. The survival analysis was performed for all codons whose translational efficiency was significantly different in more than 5 cancer types in the one direction with respect to the other [Abs(UP-DOWN)>5], and correspondingly corrected for multiple comparisons using FDR. Refer to Supplementary Table 3 for full cancer type names.

Promoter methylation and gene copy number regulate tRNA expression

Aberrant translational efficiencies in cancer are partially caused by the differential expression of tRNA genes (Figure S2). To determine the underlying mechanisms driving changes in expression, we retrieve the methylation and copy number alteration (CNA) data from TCGA samples, as a possible means for tRNA gene regulation. While CNA information cover 84% of tRNA genes, the 450K-BeadChip methylation arrays used in TCGA are mostly centered on the coding genome (Bibikova et al., 2011) and yield a coverage of only 37%.

In order to make the gene-based data comparable with the measured isoacceptor-based tRNA expression, we average methylation and CNA levels over all genes within the same isoacceptor family, at the cost of losing resolution. For each isoacceptor and each cancer type, we finally fit a Multiple Linear Regression to determine how are promoter methylation and CNA affecting tRNA expression (Figure 5A, Table S10). Among all models, the significant coefficients for methylation and CNA are significantly negative and positive, respectively. Despite the limited explained variance of the models (average R^2=0.023), such results indicate that promoter methylation contributes to inhibition of tRNA gene expression, while an increase in the gene copy number enhances tRNA expression.

Given the association of the codon ProCCA with cancer prognosis (Figure 4C), we explore the expression pattern of tRNAPro in TCGA. In particular, tRNAPro^{AGG}, which recognizes the codon ProCCA, is overexpressed in 8 out of 9 cancer types (Figure S2A). To get a more accurate picture of the tRNA gene methylation levels, we also analyze recently published bisulfite sequencing data (39), which, for 47 samples among nine cancer types, improved the coverage of tRNA genes up to an average of 81%. In total, tRNAPro^{AGG} genes stand among the most duplicated and least methylated proline isoacceptors in cancer (Figure S3A-B), in particular at the chr6.tRNA12 and chr16.tRNA12 genes (Figure 5B). Furthermore, tRNAPro^{AGG} gene duplications occur most frequently in kidney cancers (Figure S3C). On the other hand, although the other CCA-decoding tRNAPro^{TGG} is not differentially expressed in cancer (Figure S2), its genes are as similarly methylated and duplicated as tRNAPro^{AGG} (Figure 5B, Figure S3).

In short, promoter methylation and CNA appear as two possible regulatory mechanisms of tRNA expression in cancer, which suggests that similar mechanisms that control the Pol-II-mediated RNAs might also regulate the expression of Pol-III non-coding transcriptome, such as tRNA genes. However, more accurate and high-throughput data on the methylation and CNA of the non-coding genome together with gene-based tRNA quantifications are needed to make stronger associations.

Figure 5. Promoter methylation and gene copy number regulate tRNA expression. (A) A Multiple Linear Regression (MLR) between square-root-normalized tRNA expression and the average promoter methylation (450K BeadChip array) and gene copy number at the isoacceptor level. Among all MLRs for each isoacceptor and each cancer type separately,

the dots show the FDR-normalized significant coefficients based on their corresponding t-statistic p-value, and red/blue show whether they are negative/positive respectively. The p-value corresponds to a two-tailed binomial test between n_{pos} and n_{neg}. (B) Differential promoter methylation (bisulfite sequencing) between healthy and tumor samples of genes expressing proline tRNAs, as measured by $\Delta\%Me=(\%Me_{Tumor}-\%Me_{Healthy})$. Refer to Supplementary Table 3 for full cancer type names.

DISCUSSION

In this study, we use a systems biology approach to interrogate the multi-omics TCGA dataset under the perspective of translational efficiencies. We therefore first validate the suitability of small RNA-seq data in reproducing conventional tRNA-seq quantifications based on a gold standard set of five tissue-wide human cell lines. In fact, knowing that small RNA-seq datasets have a limited tRNA coverage and tend to be biased towards tRNA fragments and unmodified tRNAs (18,40), we extend and apply a computational pipeline for accurate mapping of tRNA reads (16). As a result, we obtain reproducible and informative quantifications of all isoacceptors in our gold standard cell lines as well as in thousands of samples across 23 cancer types of TCGA, exceeding the quality of similarly published data (19,20).

From these quantifications, we then elucidate their effect on the translational efficiency by defining the RTE, for Relative Translation Efficiency, which is a balance between the tRNA supply and the codon demand. Although a more accurate RTE would have determined the supply and demand based on the aminoacylated portion of tRNAs (41) and the ribosome-bound mRNAs (42) respectively, we approximate such measures by our tRNA quantifications and the publicly-available mRNA-seq data of TCGA. In agreement with current studies showing that a dynamic codon usage need to compete for a limited tRNA pool (28,29), we demonstrate that RTE is better measure of codon optimality than previously published metrics such as the tAI (32,33). However, far from explaining the translation process, the still low but significant correlations of protein-RTE in human, in contrast to unicellular organisms, suggest that protein expression is also dependent on other layers of regulation, such as transcriptional and post-transcriptional machineries, translation initiation, epigenetic modifications of DNA and RNAs, or protein degradation mechanisms (43).

On the level of translational efficiency, in agreement with previous studies (4,36), we detect that the proliferative state is the major determinant of RTE differences both across healthy tissues and in cancer. Moreover, in contrast to recent work challenging the tissue-specificity of codon-anticodon co-adaptation in human (29,43), our data here support the idea that tissue-specific RTEw have functional implications on the tissue phenotype (e.g. in determining neural differentiation in brain, or inducing abnormal proliferation in cancer). Furthermore, we observe a pattern of proliferative nnA/T versus differentiative nnC/G codons. Based on ribosome profiling experiments of pluripotency changes in embryonic stem cells (34), this could be attributed to the slower translation in differentiated cells of codons decoded by tRNAs that require adenosine-to-inosine modification at the wobble-base pairing position. In particular, we detect the ProCCA codon to be significantly more favored in proliferative cells and leading to poor cancer prognosis in kidney carcinomas, specifically driven by an overexpression of tRNAPro^{AGG} in cancer. Proline limitation in clear renal cell carcinoma has indeed been shown to mostly compromise tRNAPro^{AGG} aminoacylation, leading to slower proline translation and reduced tumor growth (38). Furthermore, in support of our approach for isoacceptor quantification and translational efficiency, similar studies of tRNA levels in TCGA have controversially claimed an opposite prognostic value for the ProCCA codon in clear renal cell carcinoma (19,20).

In an effort to elucidate the mechanisms regulating the expression of tRNAs, we observe that the tRNA gene copy number and their DNA methylation state have a positive and inhibitory effect on tRNA expression, respectively. In this context, DNA methylation has previously been linked to the silencing of type II genes (such as tRNAs) of the Pol-III transcriptome (44). Here we specifically propose a role for DNA methylation in regulating the overexpression of tRNAPro^{AGG} in cancer. In terms of the copy number alterations, it is not surprising to detect tRNA gene duplications in tumors, but the functional role in disease of different isodecoder genes that share the same anticodon is still a matter of debate (45). With the advent of more accurate and high-throughput multi-omics datasets, our knowledge on the underlying mechanisms controlling tRNA expression, degradation, and the effect of their modifications will be further expanded (8,9). Recent studies in TCGA have actually observed an upregulation of tRNA-modifying enzymes, as well as proposed a link of tRNA-derived fragments (tRF) to proliferation (19,46).

CONCLUSIONS

This is the first high-throughput study of codon-anticodon translational efficiency over thousands of samples comprising multiple tissues and disease. We therefore demonstrate a functional role for the proliferation-driven tRNA expression differences in determining a tissue-specific phenotype, both in physiological and pathological conditions. In the future, we expect to validate the effect of such differential translational efficiency by integrating perturbation□based data and including additional gene expression regulatory layers such as tRNA modifications.

METHODS

Cell lines

The cell lines included in this study are HeLa, HEK293, HCT116, MDA-MB-231 and fibroblast BJ/hTERT. Cells were maintained at 37 °C in a humidified atmosphere at 5% CO2 in DMEM 4.5g/L Glucose with UltraGlutamine media supplemented with 10% of FBS and 1% penicillin/streptomycin.

RNA extraction

Cells were grown in 60mm dishes for 48h. Total RNA from HeLa, HEK293, HCT116, MDA-MB-231 and fibroblast BJ/hTERT was extracted using the miRNeasy Mini kit. Independent replicates where grown and RNA was extracted on different days. 20 µg of total RNA was treated following either the protocol of Hydro-tRNAseq (12) or generic small RNA-seq.

Hydro-tRNA sequencing

Total RNA was resolved on 15% Novex TBE urea gels and size-selected for 60-100 nt fragments. The recovered material was then alkaline hydrolyzed (10mM sodium carbonate and 10mM sodium bicarbonate) for 10 minutes at 60ºC. The resulting RNA was de-phosphorylated with Antarctic Phosphatase (New England Biolabs) at 37ºC for 1 hour. De-phosphorylated RNA was purified with an RNeasy MinElute spin column and re-phosphorylated with Polynucleotide Kinase (NEB). PNK-treated tRNAs were purified with an RNeasy MinElute spin column and, similar to small RNA-seq library preparation, adaptor-ligated, reverse-transcribed and PCR-amplified for 14 cycles. The resulting cDNA was purified using a QIAQuick PCR Purification Kit and sequenced on Illumina HiSeq 2500 platform in 50bp paired-end format.

From all five cell lines, the isoacceptor abundances of MDA-MB-231 yielded a median of 3-5 times higher standard deviation than the other Hydro-tRNAseq quantifications (Table S2), thus suggesting some technical problem with this cell line. In consequence, this cell line was excluded from any further analysis.

Small RNA sequencing

Total RNA was directly adaptor-ligated, reverse-transcribed and PCR-amplified for 12 cycles. The resulting cDNA was purified using a QIAQuick PCR Purification Kit and sequenced on Illumina HiSeq 2500 platform in 50bp single-end format.

Computational Analysis

tRNA quantification and modification calling

In both Hydro-tRNAseq and small RNA-seq FASTQ files, sequencing adapters were trimmed using BBDuk from the BBMap toolkit [v38.22] (https://sourceforge.net/projects/bbmap): k-mer=10 (allowing 8 at the end of the read), Hamming distance=1, length=10-50bp, Phred>25. Using the human reference genome GRCh38 (Genome Reference Consortium Human Reference 38, GCA_000001405.15), a total of 856 nuclear tRNAs and 21 mitochondrial tRNAs were annotated with tRNAscan-SE [v2.0] (47).

Trimmed FASTQ files were then mapped using a specific pipeline for tRNAs (Figure 1A) (16). Summarizing, an artificial genome is first generated by masking all annotated tRNA genes and adding pre-tRNAs (i.e. tRNA genes with 3' and 5' genomic flanking regions) as extra chromosomes. Upon mapping to this artificial genome with Segemehl [v0.3.1] (48), reads that map to the tRNA-masked chromosomes or to the tRNA flanking regions are filtered out in order to remove non-tRNA reads and unmature-tRNA reads respectively.

After this first mapping step, a second library is generated by adding 3' CCA tails and removing introns from tRNA genes. All 100% identical sequences of this so-called *mature* tRNAs are clustered to avoid redundancy. Next, the subset of filtered reads from the first mapping is aligned against the clustered mature tRNAs using Segemehl [v0.3.1] (48). Mapped reads are then realigned with GATK IndelRealigner [v3.8] (49) to reduce the number of mismatching bases across all reads.

For quantification, isoacceptors were quantified as reads per million (RPM). In order to increase the coverage for anticodon-level quantification, we consider all reads that map unambiguously to a certain isoacceptor, even though they ambiguously map to different isodecoders (i.e. tRNA genes that differ in their sequence but share the same anticodon). Ambiguous reads mapping to genes of different isoacceptors were discarded.

Regarding modification site calling, we only considered gene-level uniquely mapped reads, as described to be optimal in Hoffmann et al. (16). As in their pipeline, in order to distinguish mapping or sequencing errors from true misincorporation sites, we use GATK UnifiedGenotyper [v3.8] (49).

Relative Codon Usage (RCU) and Relative Anticodon Abundance (RAA)

The RCU/RAA is defined as the contribution of a certain codon/anticodon to the amino acid it belongs to. The RCU of all synonymous codons and the RAA of all anticodons recognizing synonymous codons therefore sum up to 1.

$$RCU = \frac{x_C}{\sum_{i \in C_{aa}} x_i} \qquad RAA = \frac{x_A}{\sum_{i \in A_{aa}} x_i}$$

where x_C/x_Arefers to the abundance of the codon/anticodon C/A, and C_{aa} is the set of all synonymous codons, as well as A_{aa} is the set of all anticodons that decode synonymous codons.

tRNA Adaptation Index (tAI)

As described by dos Reis et al. (2003, 2004), the tAI weights every codon based on the wobble-base codon-anticodon interaction rules. Let c be a codon, then the decoding weight is a weighted sum of the square-root-normalized tRNA abundances $tRNA_{cj}$ for all tRNA isoacceptors j that bind with affinity $(1 - s_{cj})$ given the wobble-base pairing rules n_c. However, while dos Reis et al. (2004) assumes that highly expressed genes are codon-optimized, here we use the non-optimized s-values to avoid a circularity in our reasoning:

$$s = [0, 0, 0, 0, 0.5, 0.5, 0.75, 0.5, 0.5]$$

$$w_c = \sum_{j=1}^{n_c} (1 - s_{cj}) tRNA_{cj}$$

And therefore the tAI of a certain protein is the product of weights of each codon i_k at the triplet position k throughout the full gene length l_g, and normalized by the length.

$$tAI = (\prod_{k=1}^{l_g} w_{i_k})^{1/l_g}$$

Relative tRNA Adaptation Index (RtAI)

For comparison with the RTE (Figure 3A), an amino-acid-normalized tAI measure is defined by dividing each tAI weight by the maximum weight among all codons within each amino acid family.

$$Rw_c = \frac{w_c}{max_{i\in c_{aa}}(w_i)}$$

And therefore the RtAI of a certain protein is the product of weights Rw of each codon i_k at the triplet position k throughout the full gene length l_g, and normalized by the length.

$$RtAI = (\prod_{k=1}^{l_g} Rw_{i_k})^{1/l_g}$$

Relative Translation Efficiency (RTE)

The RTE aims to consider not only tRNA abundances, but also the codon usage demand. In doing so, it constitutes a global measure of translation control, since the efficiency of a certain codon depends both on its complementary anticodon abundance as well as the demand for such anticodon by other transcripts. This global control has been indeed established to play an important role in defining optimal translation programs (28).

The definition of the RTE is based on similar previously published metrics (26,27), which consists of a ratio between the anticodon supply and demand. On the one hand, the anticodon supply is defined as the relative tAI weights Rw (see previous section). On the other, the anticodon demand is estimated from the codon usage at the transcriptome level. It is computed as the frequency of each codon in a transcript weighted by the corresponding transcript expression, and finally summing up over all transcripts. Let c be a codon, then the codon usage is a weighted sum of the counts of codon c_i in gene j weighted by the mRNA-seq abundance $mRNA_j$ for all genes in the genome g:

$$CU_c = \sum_{j=1}^{g} c_{ij} mRNA_j$$

Similarly to the supply, the anticodon demand is then normalized within each amino acid family:

$$D_c = \frac{CU_c}{max_{i \in c_{aa}}(CU_i)}$$

Finally, the RTE weights (RTEw) are defined as the ratio between the codon supply S_c and demand D_c:

$$RTEw_c = \frac{S_c}{D_c}$$

And therefore the RTE of a certain protein is the product of weights $RTEw$ of each codon i_k at the triplet position k throughout the full gene length l_g, and normalized by the length.

$$RTE = (\prod_{k=1}^{l_g} RTEw_{i_k})^{1/l_g}$$

Gene Set Enrichment Analysis (GSEA)

We analyzed the enrichment of gene sets of the GO Biological Process Ontology using the GSEA algorithm (50). The score used to generate the ranked list input is specified in the text for each analysis.

Survival Analysis

To analyze how translational efficiency of a certain codon (RTEw) can affect the survival probability in cancer, patients of a certain cancer type are divided in two groups of low/high RTEw, which correspond to the patients having the top and bottom 40% RTEw. The Kaplan-Meier curves are then computed to estimate the survival probability of each group along time.

tRNA methylation and copy number

For consistency with the current version of publicly available and pre-processed 450k DNA methylation and SNP6 segmented copy number alteration (CNA) data from firebrowse, we used the human reference genome GRCh37/hg19 (Genome Reference Consortium Human Reference 37, GCA_000001405.1) in this analysis. The coordinates of all nuclear tRNA genes were obtained using tRNAscan-SE [v2.0] (47).

Regarding DNA methylation, we computed the average beta value of each tRNA gene from 1.5kb upstream of the transcription start site (1500TSS) until the end of the gene. For CNA, we retrieved the segmented data of precomputed $log_2(CN) - 1$ from firebrowse and extracted the corresponding value for the genomic coordinates containing the tRNA genes. Whenever the tRNA genes was located between two segments, the weighted average in function of the gene overlap with each segment was computed.

Bisulfite sequencing methylation

As 1500TSS methylation of tRNA genes lead to an average coverage of only 37% genes, we also analyzed the recently published bisulfite sequencing data of 47 samples across nine cancer types (Table S3) (39). After retrieving the datasets from the GDC legacy archive, given the higher resolution of bisulfite sequencing data, we restricted the computation of the average promoter methylation of tRNA genes to the GRCh37/hg19 genomic coordinates containing the tRNA genes, since the promoter region of Pol-III-genes is intragenic.

Multiple Linear Regression (MLR)

We fitted a Multiple Linear Regression (MLR) between the square-root-normalized tRNA expression (dependent variable) and the promoter methylation and gene copy number (independent variables). To make all three layers of information comparable, we considered only samples for which all data was available and performed the regression at the isoacceptor level, thus averaging the methylation and CNA data over all tRNA genes that shared the same anticodon.

$$EXP = \beta_0 + \beta_{Me}Me + \beta_{CNA}CNA$$

We fitted the model parameters for all 64 isoacceptors and 22 cancer types, leading to 22x64=1408 MLRs, among which only significant coefficients (FDR-corrected t-statistic p-value < 0.05) were considered in downstream analyses.

Statistical Analysis

For hypothesis testing, an unpaired two-tailed Wilcoxon rank-sum test was performed, unless stated otherwise. All details of the statistical analyses can be found in the Results section. We used a significance value of 0.05. In differential expression analyses, a False Discovery Rate correction was used to account for multiple testing.

ABBREVIATIONS

TCGA: The Cancer Genome Atlas

PCA: Principal Component Analysis

GSEA: Gene Set Enrichment Analysis

RAA: Relative Anticodon Abundance

RTE: Relative Translation Efficiency

RTEw: RTE weights

tAI: tRNA Adaptation Index

RtAI: Relative tAI

CNA: Copy Number Alteration

BLCA: Bladder Urothelial Carcinoma

BRCA: Breast invasive carcinoma

CESC: Cervical squamous cell carcinoma and endocervical adenocarcinoma

CHOL: Cholangiocarcinoma

COAD: Colon adenocarcinoma

ESCA: Esophageal carcinoma

GBM: Glioblastoma multiforme

HNSC: Head and Neck squamous cell carcinoma

KICH: Kidney Chromophobe

KIRC: Kidney renal clear cell carcinoma

KIRP: Kidney renal papillary cell carcinoma

LIHC: Liver hepatocellular carcinoma

LUAD: Lung adenocarcinoma

LUSC: Lung squamous cell carcinoma

PAAD: Pancreatic adenocarcinoma

PCPG: Pheochromocytoma and Paraganglioma

PRAD: Prostate adenocarcinoma

READ: Rectum adenocarcinoma

SKCM: Skin Cutaneous Melanoma

STAD: Stomach adenocarcinoma

THCA: Thyroid carcinoma

THYM: Thymoma

UCEC: Uterine Corpus Endometrial Carcinoma

DECLARATIONS

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

CONSENT FOR PUBLICATION

Not applicable.

AVAILABILITY OF DATA AND MATERIALS

HEK293 Small RNA-seq datasets

The four HEK293 datasets were downloaded from the NCBI Sequence Read Archive (SRA): SRR1304304, ERR705692, ERR705691, SRR2060090.

The Cancer Genome Atlas

Raw small RNA-sequencing data in BAM format were retrieved from the GDC legacy archive after obtaining the necessary permissions from dbGaP, comprising all healthy samples (NT, solid tissue normal) and their primary tumor (PT) counterparts, which consists of 23 cancer types (BRCA, PRAD, KICH, KIRP, KIRC, LUAD, LUSC, HNSC, UCEC, CESC, LIHC, CHOL, THCA, COAD, READ, ESCA, STAD, BLCA, PAAD, THYM, SKCM, PCPG, GBM). For samples for which more than one BAM was available, all files were downloaded. BAM files were converted to FASTQ using SAMtools [v1.3.1] (51). We retrieved publicly available and pre-processed mRNA-seq gene expression, 450k DNA methylation, and SNP6 segmented copy number alteration (CNA) from firebrowse. As for proteomics, preprocessed protein assembly data and protein relative abundance were obtained from CPTAC for TCGA samples including BRCA, COAD and READ.

Coding sequences

The coding sequences of *Homo sapiens* from RefSeq were downloaded from the Codon/Codon Pair Usage Tables (CoCoPUTs) project release as of February 6, 2019 (52,53).

GO gene sets

Gene sets derived from the GO Biological Process Ontology were downloaded from the Molecular Signatures Database [v6.2] (MSigDB) as a GMT file (50,54).

Generated data and code

The code used in this study is available at GitHub [https://github.com/hexavier/tRNA_TCGA; https://github.com/hexavier/tRNA_mapping], and the generated datasets are publicly accessible at Synapse (www.synapse.org/tRNA_TCGA, syn20640275). Hydro-tRNA and small RNA sequencing data of all five cell lines has been made available at the Gene Expression Omnibus (GEO): GSE137834. Hydro-tRNAseq data from HEK293 and HeLa has been previously published (36) and deposited at ArrayExpress under accession number E-MTAB-8144.

COMPETING INTERESTS

The authors declare that they have no competing interests.

FUNDING

We acknowledge the support of the Spanish Ministry of Economy and Competitiveness, ‘Centro de Excelencia Severo Ochoa’, the CERCA Programme / Generalitat de Catalunya, and the Spanish Ministry of Economy, Industry and Competitiveness (MEIC) to the EMBL partnership. The work of X.H. has been supported by a PhD fellowship from the Fundación Ramón Areces.

AUTHORS' CONTRIBUTIONS

Conceptualization, X.H., M.H.S. and L.S.; Methodology, X.H., H.B., M.H.S., L.S.; Software, X.H.; Investigation, H.B., X.H.; Validation, X.H., M.H.S.; Formal analysis, X.H., M.H.S.; Writing-Original Draft, X.H.; Writing-Review & Editing, X.H., H.B., M.H.S., L.S.; Visualisation: X.H., M.H.S.; Funding Acquisition, L.S.; Supervision, M.H.S. and L.S.

ACKNOWLEDGEMENTS

We thank Eva Maria Novoa Pardo, Samuel Miravet-Verde, and Marc Weber for stimulating and critical discussions. We thank the CRG Genomics Unit for assistance with RNA sequencing services. The results published here are in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

REFERENCES

1. Hanson G, Coller J. Codon optimality, bias and usage in translation and mRNA decay. Nat Rev Mol Cell Biol. 2017 Oct 11;19(1):20–30.
2. Supek F. The Code of Silence: Widespread Associations Between Synonymous Codon Biases and Gene Function. J Mol Evol. 2016 Jan;82(1):65–73.
3. Najafabadi HS, Goodarzi H, Salavati R. Universal function-specificity of codon usage. Nucleic Acids Res. 2009 Nov;37(21):7014–23.
4. Gingold H, Tehler D, Christoffersen NR, Nielsen MM, Asmar F, Kooistra SM, et al. A Dual Program for Translation Regulation in Cellular Proliferation and Differentiation. Cell. 2014 Sep;158(6):1281–92.
5. Pouyet F, Mouchiroud D, Duret L, Sémon M. Recombination, meiotic expression and human codon usage. eLife [Internet]. 2017 Aug 15 [cited 2018 Nov 7];6. Available from: https://elifesciences.org/articles/27344
6. Chan PP, Lowe TM. GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. Nucleic Acids Res. 2016 Jan 4;44(D1):D184-189.
7. Schimmel P. The emerging complexity of the tRNA world: mammalian tRNAs beyond protein synthesis. Nat Rev Mol Cell Biol. 2018 Jan;19(1):45–58.
8. Pan T. Modifications and functional genomics of human transfer RNA. Cell Res. 2018 Apr;28(4):395.
9. Rak R, Dahan O, Pilpel Y. Repertoires of tRNAs: The Couplers of Genomics and Proteomics. Annu Rev Cell Dev Biol. 2018 Oct 6;34(1):239–64.
10. Dittmar KA, Goodenbour JM, Pan T. Tissue-Specific Differences in Human Transfer RNA Expression. PLOS Genet. 2006 Dec 22;2(12):e221.
11. Waldman YY, Tuller T, Shlomi T, Sharan R, Ruppin E. Translation efficiency in humans: tissue specificity, global optimization and differences between developmental stages. Nucleic Acids Res. 2010 May 1;38(9):2964–74.
12. Gogakos T, Brown M, Garzia A, Meyer C, Hafner M, Tuschl T. Characterizing Expression and Processing of Precursor and Mature Human tRNAs by Hydro-tRNAseq and PAR-CLIP. Cell Rep. 2017 Aug;20(6):1463–75.

13. Zheng G, Qin Y, Clark WC, Dai Q, Yi C, He C, et al. Efficient and quantitative high-throughput tRNA sequencing. Nat Methods. 2015 Sep;12(9):835–7.
14. Guo Y, Bosompem A, Mohan S, Erdogan B, Ye F, Vickers KC, et al. Transfer RNA detection by small RNA deep sequencing and disease association with myelodysplastic syndromes. BMC Genomics. 2015 Sep 24;16(1):727.
15. Guo Y, Xiong Y, Sheng Q, Zhao S, Wattacheril J, Flynn CR. A micro-RNA expression-signature for human NAFLD progression. J Gastroenterol. 2016 Oct;51(10):1022–30.
16. Hoffmann A, Fallmann J, Vilardo E, Mörl M, Stadler PF, Amman F. Accurate mapping of tRNA reads. Bioinformatics. 2018 Apr 1;34(7):1116–24.
17. Pundhir S, Gorodkin J. Differential and coherent processing patterns from small RNAs. Sci Rep [Internet]. 2015 Jul 13 [cited 2019 Jun 21];5. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4499813/
18. Torres AG, Piñeyro D, Rodríguez-Escribà M, Camacho N, Reina O, Saint-Léger A, et al. Inosine modifications in human tRNAs are incorporated at the precursor tRNA level. Nucleic Acids Res. 2015 May 26;43(10):5145–57.
19. Zhang Z, Ye Y, Gong J, Ruan H, Liu C-J, Xiang Y, et al. Global analysis of tRNA and translation factor expression reveals a dynamic landscape of translational regulation in human cancers. Commun Biol. 2018 Dec 21;1(1):234.
20. Zhang Z, Ruan H, Liu C-J, Ye Y, Gong J, Diao L, et al. tRic: a user-friendly data portal to explore the expression landscape of tRNAs in human cancers. RNA Biol. 2019 Aug 25;
21. Mattijssen S, Arimbasseri AG, Iben JR, Gaidamakov S, Lee J, Hafner M, et al. LARP4 mRNA codon-tRNA match contributes to LARP4 activity for ribosomal protein mRNA poly(A) tail length protection. eLife [Internet]. 2017 Sep 12 [cited 2019 Feb 27];6. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5626478/
22. Chu A, Robertson G, Brooks D, Mungall AJ, Birol I, Coope R, et al. Large-scale profiling of microRNAs for The Cancer Genome Atlas. Nucleic Acids Res. 2016 Jan 8;44(1):e3.
23. Flores O, Kennedy EM, Skalsky RL, Cullen BR. Differential RISC association of endogenous human microRNAs predicts their inhibitory potential. Nucleic Acids Res. 2014 Apr;42(7):4629–39.
24. Mefferd AL, Kornepati AVR, Bogerd HP, Kennedy EM, Cullen BR. Expression of CRISPR/Cas single guide RNAs using small tRNA promoters. RNA N Y N. 2015 Sep;21(9):1683–9.
25. Scholzen T, Gerdes J. The Ki-67 protein: From the known and the unknown. J Cell Physiol. 2000;182(3):311–22.
26. Gingold H, Dahan O, Pilpel Y. Dynamic changes in translational efficiency are deduced from codon usage of the transcriptome. Nucleic Acids Res. 2012 Nov 1;40(20):10053–63.
27. Pechmann S, Frydman J. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. Nat Struct Mol Biol. 2013 Feb;20(2):237–43.
28. Frumkin I, Lajoie MJ, Gregg CJ, Hornung G, Church GM, Pilpel Y. Codon usage of highly expressed genes affects proteome-wide translation efficiency. Proc Natl Acad Sci U S A. 2018 22;115(21):E4940–9.
29. Eraslan B, Wang D, Gusic M, Prokisch H, Hallström BM, Uhlén M, et al. Quantification and discovery of sequence determinants of protein□per□mRNA amount in 29 human tissues. Mol Syst Biol. 2019 Feb 1;15(2):e8513.
30. Mertins P, Mani DR, Ruggles KV, Gillette MA, Clauser KR, Wang P, et al. Proteogenomics connects somatic mutations to signalling in breast cancer. Nature. 2016 Jun;534(7605):55–62.
31. Slebos RJC, Wang X, Wang X, Zhang B, Tabb DL, Liebler DC. Proteomic analysis of colon and rectal carcinoma using standard and customized databases. Sci Data. 2015 Jun 23;2:150022.
32. dos Reis M, Wernisch L, Savva R. Unexpected correlations between gene expression and codon usage bias from microarray data for the whole Escherichia coli K-12

genome. Nucleic Acids Res. 2003 Dec 1;31(23):6976–85.

33. dos Reis M, Savva R, Wernisch L. Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res. 2004 Jan 1;32(17):5036–44.
34. Bornelöv S, Selmi T, Flad S, Dietmann S, Frye M. Codon usage optimization in pluripotent embryonic stem cells. Genome Biol. 2019 Jun 7;20(1):119.
35. Fornasiero EF, Rizzoli SO. Pathological changes are associated with shifts in the employment of synonymous codons at the transcriptome level. BMC Genomics. 2019 Jul 9;20(1):566.
36. Benisty H, Weber M, Hernandez-Alias X, Schaefer MH, Serrano L. Proliferation specific codon usage facilitates oncogene translation. Manuscr Submitt Publ [Internet]. 2019 Jul 11 [cited 2019 Jul 24]; Available from: https://www.biorxiv.org/content/10.1101/695957v1
37. Goodarzi H, Nguyen HCB, Zhang S, Dill BD, Molina H, Tavazoie SF. Modulated Expression of Specific tRNAs Drives Gene Expression and Cancer Progression. Cell. 2016 Jun;165(6):1416–27.
38. Loayza-Puch F, Rooijers K, Buil LCM, Zijlstra J, F. Oude Vrielink J, Lopes R, et al. Tumour-specific proline vulnerability uncovered by differential ribosome codon reading. Nature. 2016 Feb;530(7591):490–4.
39. Zhou W, Dinh HQ, Ramjan Z, Weisenberger DJ, Nicolet CM, Shen H, et al. DNA methylation loss in late-replicating domains is linked to mitotic cell division. Nat Genet. 2018 Apr;50(4):591–602.
40. Torres AG, Reina O, Attolini CS-O, Pouplana LR de. Differential expression of human tRNA genes drives the abundance of tRNA-derived fragments. Proc Natl Acad Sci. 2019 Apr 23;116(17):8451–6.
41. Evans ME, Clark WC, Zheng G, Pan T. Determination of tRNA aminoacylation levels by high-throughput sequencing. Nucleic Acids Res. 2017 Aug 21;45(14):e133.
42. Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS. Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. Science. 2009 Apr 10;324(5924):218–23.
43. Rudolph KLM, Schmitt BM, Villar D, White RJ, Marioni JC, Kutter C, et al. Codon-Driven Translational Efficiency Is Stable across Diverse Mammalian Cell States. Galtier N, editor. PLOS Genet. 2016 May 11;12(5):e1006024.
44. Park J-L, Lee Y-S, Kunkeaw N, Kim S-Y, Kim I-H, Lee YS. Epigenetic regulation of noncoding RNA transcription by mammalian RNA polymerase III. Epigenomics. 2017 Jan 23;9(2):171–87.
45. Lant JT, Berg MD, Heinemann IU, Brandl CJ, O'Donoghue P. Pathways to disease from natural variations in human cytoplasmic tRNAs. J Biol Chem. 2019 Apr 5;294(14):5294–308.
46. Telonis AG, Loher P, Magee R, Pliatsika V, Londin E, Kirino Y, et al. tRNA Fragments Show Intertwining with mRNAs of Specific Repeat Content and Have Links to Disparities. Cancer Res. 2019 Jan 1;canres.0789.2019.
47. Chan PP, Lowe TM. tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences. Methods Mol Biol Clifton NJ. 2019;1962:1–14.
48. Hoffmann S, Otto C, Kurtz S, Sharma CM, Khaitovich P, Vogel J, et al. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. PLoS Comput Biol. 2009 Sep;5(9):e1000502.
49. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep;20(9):1297–303.
50. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005 Oct 25;102(43):15545–50.
51. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinforma Oxf Engl. 2009 Aug 15;25(16):2078–9.

52. Alexaki A, Kames J, Holcomb DD, Athey J, Santana-Quintero LV, Lam PVN, et al. Codon and Codon-Pair Usage Tables (CoCoPUTs): Facilitating Genetic Variation Analyses and Recombinant Gene Design. J Mol Biol. 2019 Jun 14;431(13):2434–41.
53. Athey J, Alexaki A, Osipova E, Rostovtsev A, Santana-Quintero LV, Katneni U, et al. A new and updated resource for codon usage tables. BMC Bioinformatics. 2017 Sep 2;18(1):391.
54. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database Hallmark Gene Set Collection. Cell Syst. 2015 Dec 23;1(6):417–25.
55. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinforma Oxf Engl. 2010 Mar 15;26(6):841–2.

ADDITIONAL FILES

Additional file 1 - Supplemental Methods

Format: PDF (.pdf)

List of reagents and detailed description the computational software used.

Additional file 2 - Supplemental Figures

Format: PDF (.pdf)

Figure S1. Medians of Relative Translation Efficiencies weights (RTEw) across all TCGA tissues, related to Figure 3. **Figure S2.** Differential expression of tRNAs between healthy and tumor samples across 22 cancer types, related to Figure 4. **Figure S3.** Differential methylation and copy number between healthy and tumor samples of tRNA genes, related to Figure 5.

Additional file 3 - Table S1. Correlation Matrix

Format: Comma Separated Values (.csv)

Correlations of tRNA expression of HEK293 from four small RNA-seq datasets against three conventional tRNA-seq quantifications.

Additional file 4 - Table S2. Sequencing Quality

Format: XLSX (.xlsx)

Comparison of tRNA reads coverage between small RNA-seq and hydro-tRNAseq datasets, as well as the coefficient of variance and the standard deviation between replicates.

Additional file 5 - Table S3. TCGA samples

Format: Tab Separated Values (.tsv)

Number and abbreviations of TCGA samples covering 23 cancer types.

Additional file 6 - Table S4. GSEA RAA

Format: XLSX (.xlsx)

Component features of the Principal Component Analysis of the Relative Anticodon Abundances (RAA, see Figure 2B). GSEA of the correlations of the first PCA component against all genes.

Additional file 7 - Table S5. GSEA RTE

Format: XLSX (.xlsx)

Component features of the Principal Component Analysis of the Relative Tranlational Efficiency weights (RTEw, see Figure 3B). GSEA of the correlations of the first two PCA components against all genes.

Additional file 8 - Table S6. GSEA deltaRTE

Format: XLSX (.xlsx)

GSEA of the differential RTE between extreme tissues ($\Delta RTE = RTE_{Colorectal} - RTE_{Brain}$).

Additional file 9 - Table S7. RTE survival analysis

Format: Comma Separated Values (.csv)

Association of RTEw with cancer prognosis across 22 cancer types. The survival analysis was performed for all codons whose translational efficiency was significantly different in more than 5 cancer types in the one direction with respect to the other [Abs(UP-DOWN)>5], and correspondingly corrected for multiple comparisons using FDR.

Additional file 10 - Table S8. GSEA deltaRTE KIRC

Format: XLSX (.xlsx)

GSEA of the differential RTE between cancer and tumor samples from KIRC ($\Delta RTE = RTE_{Tumor} - RTE_{Healthy}$).

Additional file 11 - Table S9. GSEA RCU ProCCA

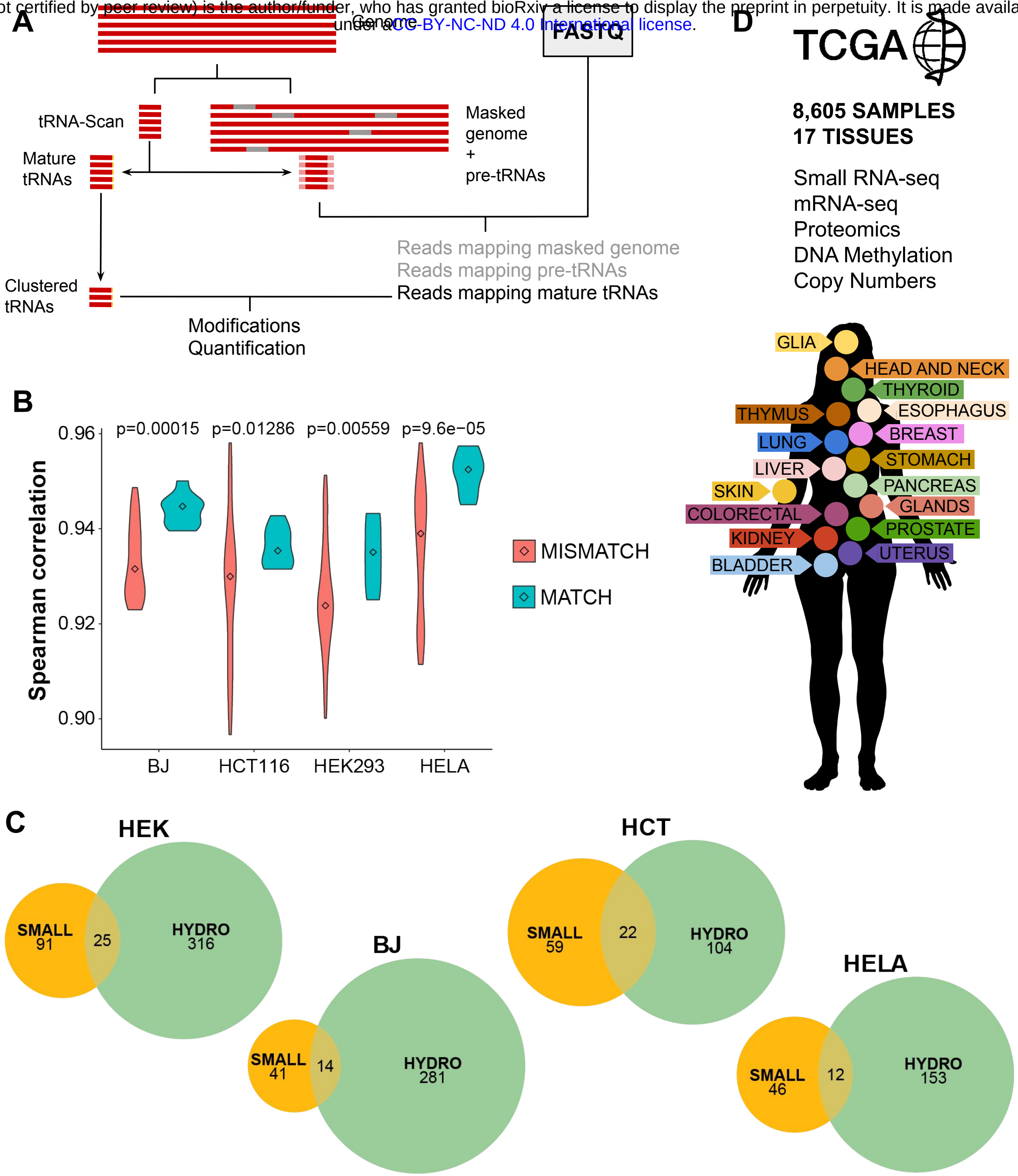
Format: XLSX (.xlsx)

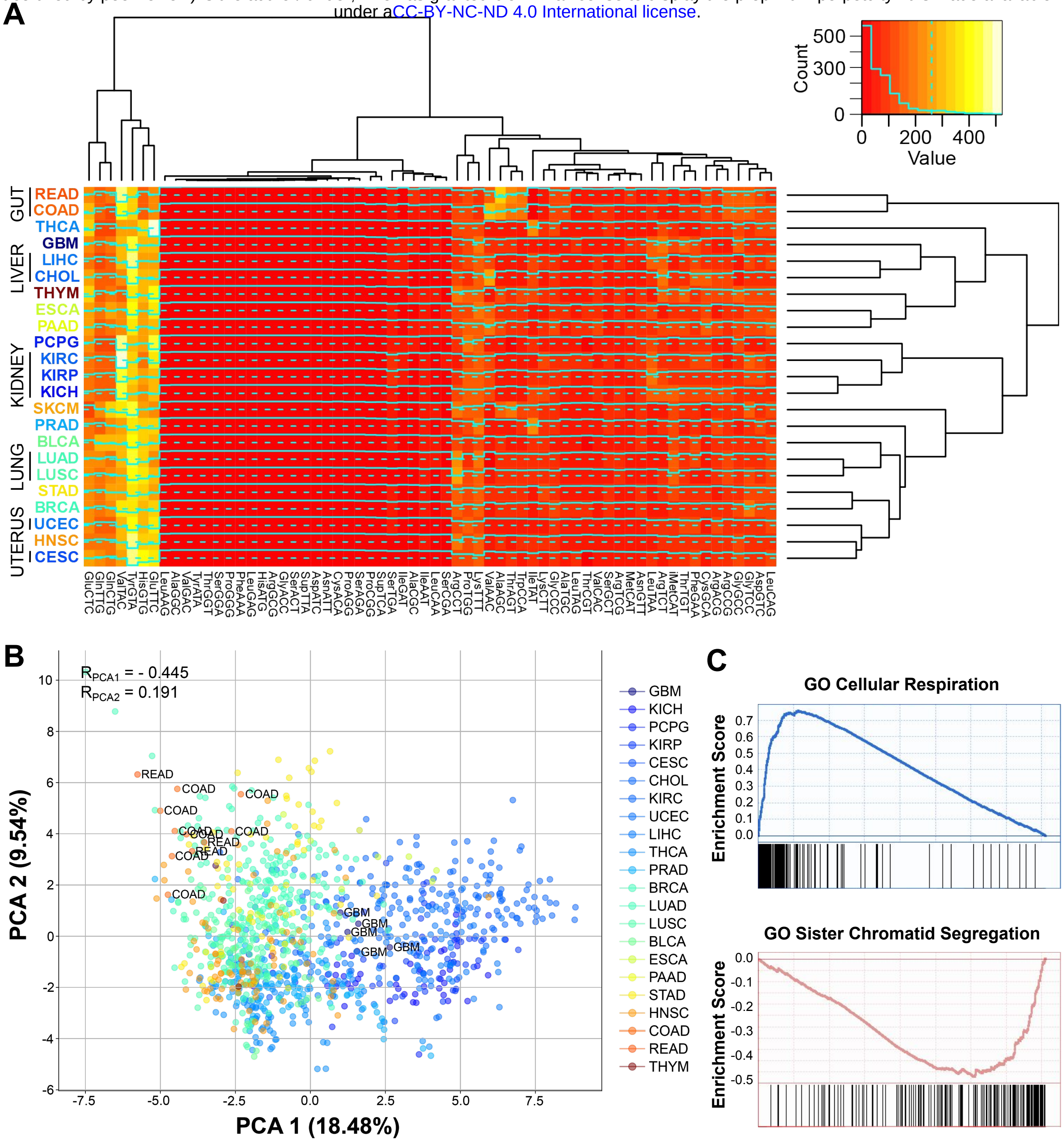
GSEA of the Relative Codon Usage (RCU) of the codon ProCCA among the whole genome.

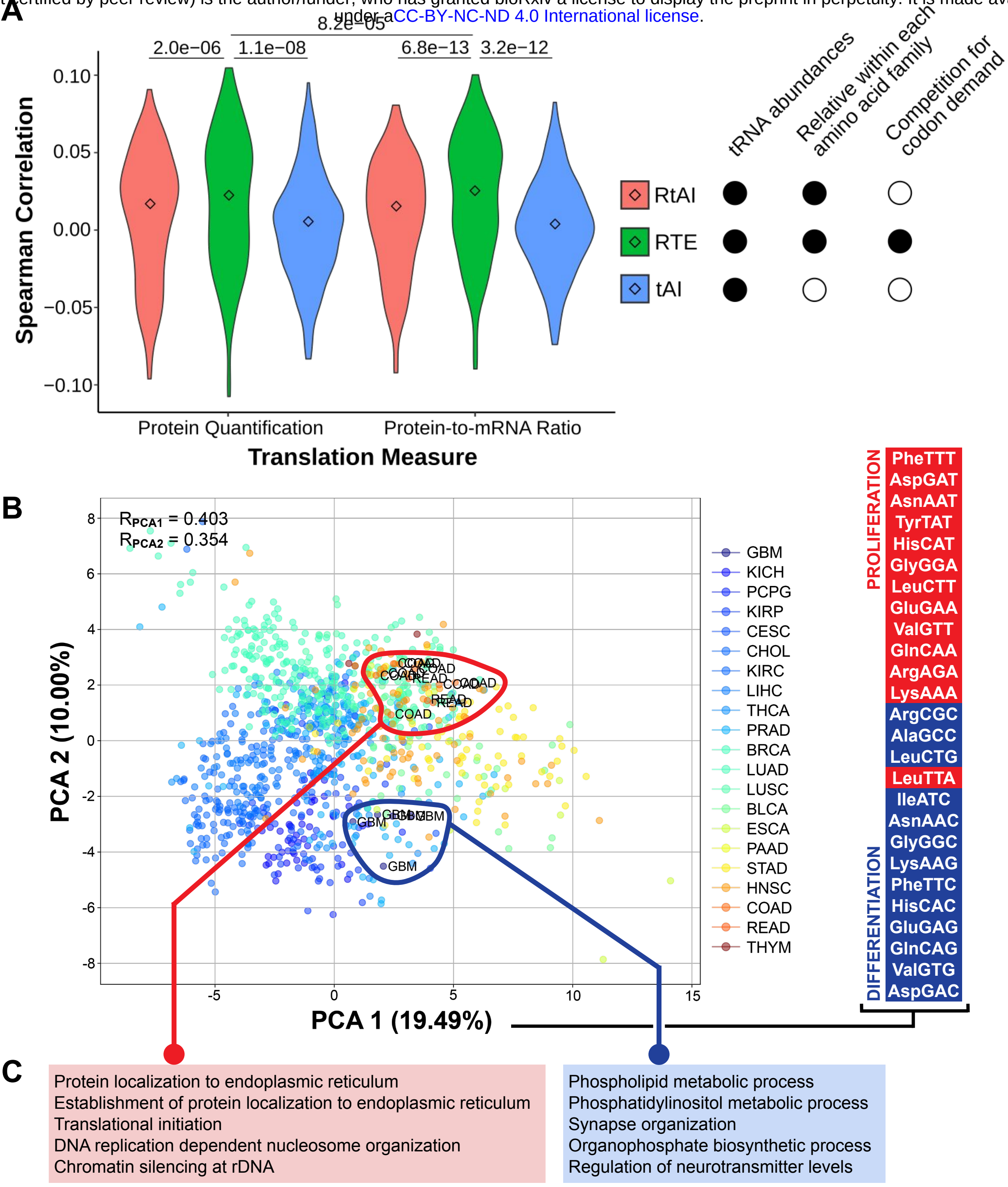
Additional file 12 - Table S10. MLR coefficients

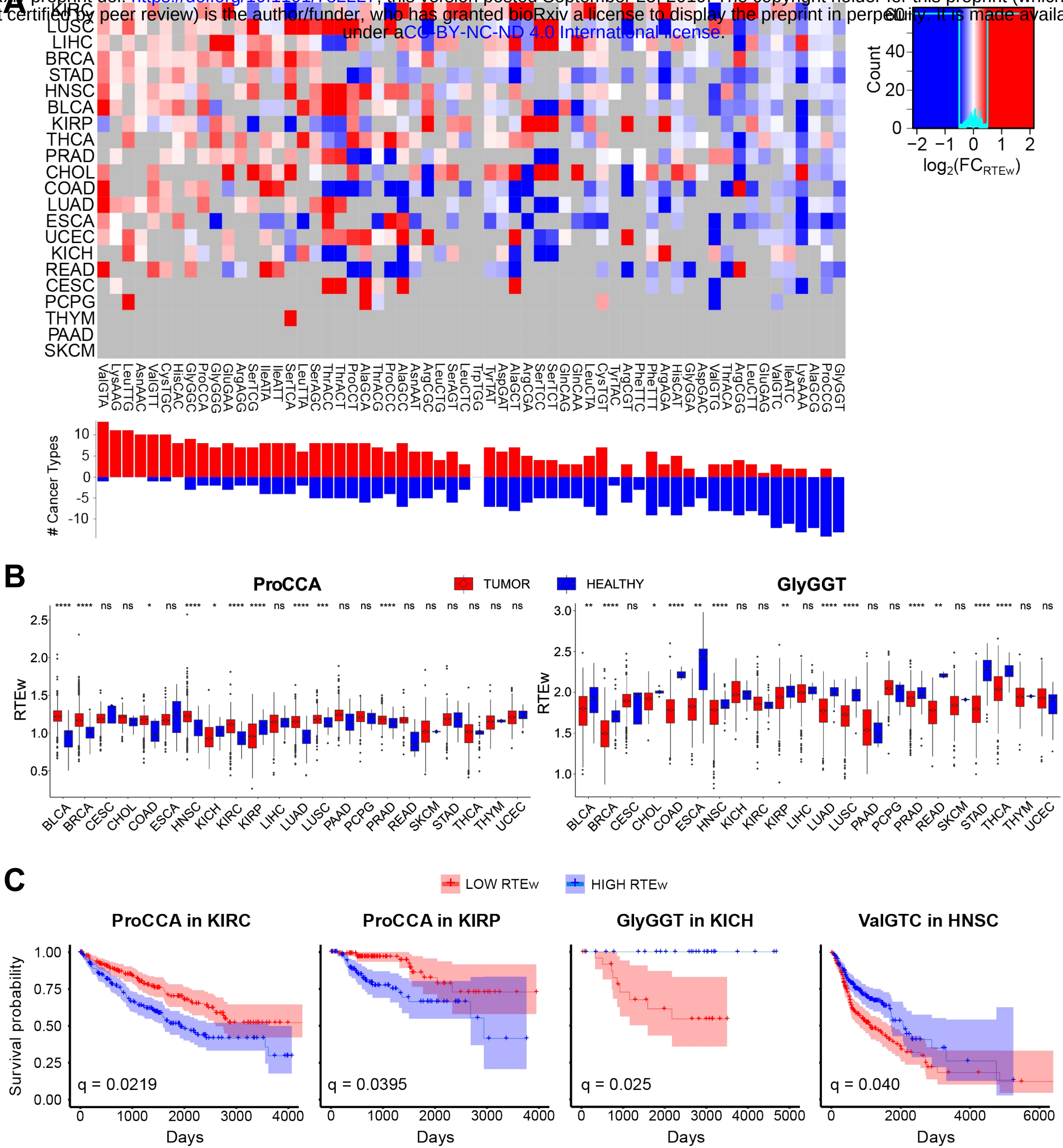
Format: XLSX (.xlsx)

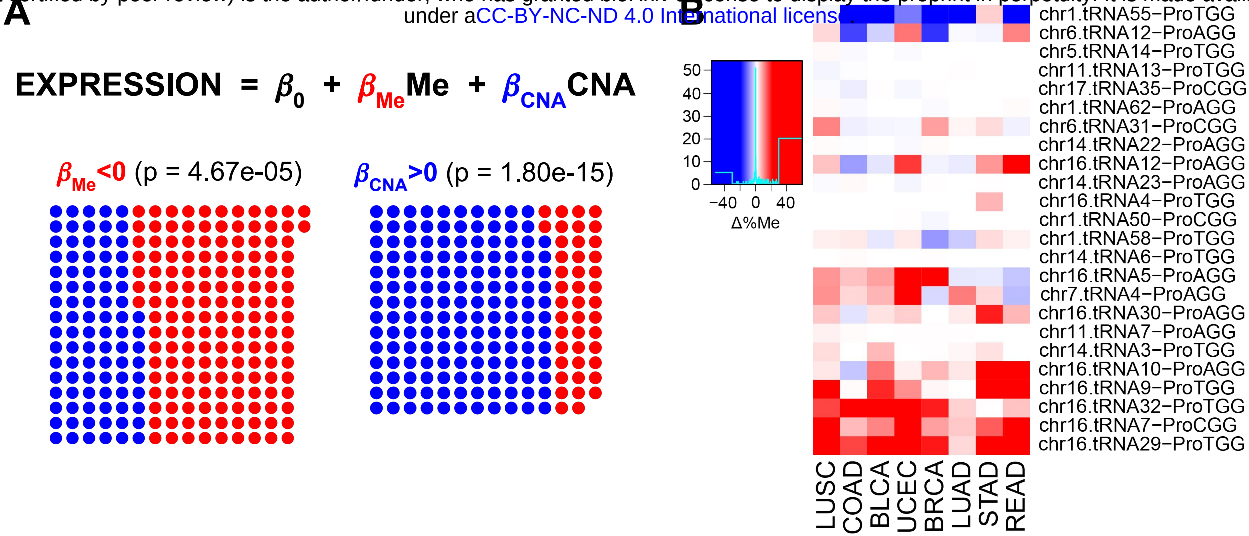
Multiple Linear Regression (MLR) between the square-root-normalized tRNA expression and the promoter methylation and gene copy number.


A
Genome
FASTQ
tRNA-Scan
Masked
genome
+
pre-tRNAs
Mature
tRNAs
Reads mapping masked genome
Reads mapping pre-tRNAs
Reads mapping mature tRNAs
Clustered
tRNAs
Modifications
Quantification
B
p=0.00015
p=0.01286
p=0.00559
p=9.6e−05
Spearman correlation
0.96
0.94
0.92
0.90
BJ
HCT116
HEK293
HELA
MISMATCH
MATCH
C
HEK
SMALL 91
25
HYDRO 316
BJ
SMALL 41
14
HYDRO 281
HCT
SMALL 59
22
HYDRO 104
HELA
SMALL 46
12
HYDRO 153
D
TCGA
8,605 SAMPLES
17 TISSUES
Small RNA-seq
mRNA-seq
Proteomics
DNA Methylation
Copy Numbers
GLIA
HEAD AND NECK
THYROID
THYMUS
ESOPHAGUS
LUNG
BREAST
LIVER
STOMACH
SKIN
PANCREAS
GLANDS
COLORECTAL
PROSTATE
KIDNEY
UTERUS
BLADDER


A
Count
500
300
0
0
200
400
Value
GUT
LIVER
KIDNEY
LUNG
UTERUS
READ
COAD
THCA
GBM
LIHC
CHOL
THYM
ESCA
PAAD
PCPG
KIRC
KIRP
KICH
SKCM
PRAD
BLCA
LUAD
LUSC
STAD
BRCA
UCEC
HNSC
CESC
GluCTC
GlnTTG
GlnCTG
ValTAC
TyrGTA
HisGTG
GluTTC
LeuAAG
AlaGGC
ValGAC
TyrATA
ThrGGT
SerGGA
ProGGG
PheAAA
LeuGAG
HisATG
ArgGCG
GlyACC
SerACT
SupTTA
AspATC
AsnATT
CysACA
ProAGG
SerAGA
ProCGG
SupTCA
SerTGA
IleGAT
AlaCGC
IleAAT
LeuCAA
SerCGA
ArgCCT
ProTGG
LysTTT
ValAAC
AlaAGC
ThrAGT
TrpCCA
IleTAT
LysCTT
GlyCCC
AlaTGC
LeuTAG
ThrCGT
ValCAC
SerGCT
ArgTCG
MetCAT
AsnGTT
LeuTAA
ArgTCT
iMetCAT
ThrTGT
PheGAA
CysGCA
ArgACG
ArgCCG
GlyGCC
GlyTCC
AspGTC
LeuCAG
B
R_{PCA1} = - 0.445
R_{PCA2} = 0.191
READ
COAD
COAD
COAD
COAD
COAD
COAD
READ
READ
COAD
COAD
GBM
GBM
GBM
GBM
GBM
PCA 2 (9.54%)
PCA 1 (18.48%)
10
8
6
4
2
0
-2
-4
-6
-7.5
-5.0
-2.5
0.0
2.5
5.0
7.5
GBM
KICH
PCPG
KIRP
CESC
CHOL
KIRC
UCEC
LIHC
THCA
PRAD
BRCA
LUAD
LUSC
BLCA
ESCA
PAAD
STAD
HNSC
COAD
READ
THYM
C
GO Cellular Respiration
Enrichment Score
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
GO Sister Chromatid Segregation
Enrichment Score
0.0
-0.1
-0.2
-0.3
-0.4
-0.5


A
8.2e−05
2.0e−06
1.1e−08
6.8e−13
3.2e−12
Spearman Correlation
0.10
0.05
0.00
−0.05
−0.10
Protein Quantification
Protein-to-mRNA Ratio
Translation Measure
tRNA abundances
Relative within each amino acid family
Competition for codon demand
RtAI ● ● ○
RTE ● ● ●
tAI ● ○ ○
B
R_PCA1 = 0.403
R_PCA2 = 0.354
PCA 2 (10.00%)
PCA 1 (19.49%)
8
6
4
2
0
-2
-4
-6
-8
-5
0
5
10
15
COAD
READ
GBM
GBM
KICH
PCPG
KIRP
CESC
CHOL
KIRC
LIHC
THCA
PRAD
BRCA
LUAD
LUSC
BLCA
ESCA
PAAD
STAD
HNSC
COAD
READ
THYM
PROLIFERATION
PheTTT
AspGAT
AsnAAT
TyrTAT
HisCAT
GlyGGA
LeuCTT
GluGAA
ValGTT
GlnCAA
ArgAGA
LysAAA
ArgCGC
AlaGCC
LeuCTG
LeuTTA
IleATC
AsnAAC
GlyGGC
LysAAG
PheTTC
HisCAC
GluGAG
GlnCAG
ValGTG
AspGAC
DIFFERENTIATION
C
Protein localization to endoplasmic reticulum
Establishment of protein localization to endoplasmic reticulum
Translational initiation
DNA replication dependent nucleosome organization
Chromatin silencing at rDNA
Phospholipid metabolic process
Phosphatidylinositol metabolic process
Synapse organization
Organophosphate biosynthetic process
Regulation of neurotransmitter levels


A
KIRC
LUSC
LIHC
BRCA
STAD
HNSC
BLCA
KIRP
THCA
PRAD
CHOL
COAD
LUAD
ESCA
UCEC
KICH
READ
CESC
PCPG
THYM
PAAD
SKCM
ValGTA LysAAG LeuTTG AsnAAC ValGTT CysTGC HisCAC GlyGGC ProCCA GlyGGG GluGAA ArgACG SerTCG IleATA IleATT SerTCA LeuTTA SerAGC ThrACC ThrACT ProCCT AlaGCA ThrACG ProCCC AlaGCC AsnAAT ArgCGC LeuCTG SerAGT LeuCTC TrpTGG TyrTAT AspGAT AlaGCT ArgCGA SerTCC SerTCT GlnCAG GlnCAA LeuCTA CysTGT TyrTAC ArgCGT PheTTC PheTTT ArgAGA HisCAT GlyGGA AspGAC ValGTG ThrACA ArgCGG LeuCTT GluGAG ValGTC IleATC LysAAA AlaGCG ProCCG GlyGGT
Count
60
40
20
0
−2 −1 0 1 2
log2(FCRTEw)
# Cancer Types
10
5
0
-5
-10
B
ProCCA
TUMOR
HEALTHY
GlyGGT
**** **** ns ns * ns **** * **** **** ns **** *** ns ns **** ns ns ns ns ns ns
RTEw
2.5
2.0
1.5
1.0
0.5
** **** ns * **** ** **** ns ns ** ns **** **** ns ns **** ** ns **** **** ns ns
3.0
2.5
2.0
1.5
1.0
BLCA BRCA CESC CHOL COAD ESCA HNSC KICH KIRC KIRP LIHC LUAD LUSC PAAD PCPG PRAD READ SKCM STAD THCA THYM UCEC
C
LOW RTEw
HIGH RTEw
ProCCA in KIRC
ProCCA in KIRP
GlyGGT in KICH
ValGTC in HNSC
Survival probability
1.00
0.75
0.50
0.25
0.00
q = 0.0219
q = 0.0395
q = 0.025
q = 0.040
0 1000 2000 3000 4000
0 1000 2000 3000 4000
0 1000 2000 3000 4000 5000
0 2000 4000 6000
Days


A
EXPRESSION = β_0 + β_{Me}Me + β_{CNA}CNA
β_{Me}<0 (p = 4.67e-05)
β_{CNA}>0 (p = 1.80e-15)
B
50
40
30
20
10
0
−40
0
40
Δ%Me
chr1.tRNA55−ProTGG
chr6.tRNA12−ProAGG
chr5.tRNA14−ProTGG
chr11.tRNA13−ProTGG
chr17.tRNA35−ProCGG
chr1.tRNA62−ProAGG
chr6.tRNA31−ProCGG
chr14.tRNA22−ProAGG
chr16.tRNA12−ProAGG
chr14.tRNA23−ProAGG
chr16.tRNA4−ProTGG
chr1.tRNA50−ProCGG
chr1.tRNA58−ProTGG
chr14.tRNA6−ProTGG
chr16.tRNA5−ProAGG
chr7.tRNA4−ProAGG
chr16.tRNA30−ProAGG
chr11.tRNA7−ProAGG
chr14.tRNA3−ProTGG
chr16.tRNA10−ProAGG
chr16.tRNA9−ProTGG
chr16.tRNA32−ProTGG
chr16.tRNA7−ProCGG
chr16.tRNA29−ProTGG
LUSC
COAD
BLCA
UCEC
BRCA
LUAD
STAD
READ